氏　　名	大田 達郎
学位(専攻分野)	博士(理学)
学 位 記 番 号	総研大乙第 260 号
学位授与の日付	平成 31年 3 月22日
学位授与の要件	学位規則第6条第2項該当
学 位 論 文 題 目	Eliminating uncertainty in the process of knowledge acquisition from the large-scale genomic data
論 文 審 査 委 員	主　査　教授 有田 正規 教授 黒川 顕 教授 木村 暁 准教授 川本 祥子 准教授 門田 幸二 東京大学大学院農学生命科学研究科

（様式３）

博士論文の要旨

氏　　名　大田　達郎

論文題目　Eliminating uncertainty in the process of knowledge acquisition from the large-scale genomic data

Genomic science has become a big data science since the advent of the high-throughput sequencing (HTSeq) technologies which produce a massive amount of nucleotide sequence data. Sequence Read Archive (SRA) is a public data repository for the HTSeq, where researchers submit the raw data from HTSeq experiments, now archives more than 4 million samples.　To extract biological knowledge from these big data of genome sequences, researchers need to use computational software to perform various kinds of data analysis.

Performing genomic data analysis is often a complicated process because many factors affect the application of data analysis software. For example, researchers need to confirm the target molecules of the sequencing experiment, the nature of the sequenced sample, the applied experimental instruments, and the used reagents to select appropriate software for data analysis. Researchers also have to understand the software operation often with many options and input parameters. Without sufficient background information and accurate operation of software, one cannot perform a proper data analysis, which results in producing the unreliable output. Thus, the precise description of the data analysis process is a key to evaluate the output of the research. However, it is not a manageable task for researchers to describe the precise process of knowledge extraction from the data without a system to support. Therefore, in this research through the case of database development from the public sequencing data, I propose the methods to describe the data analysis process to remove uncertainty.

To describe the precise information of input data for the analysis, I developed a system to integrate sample metadata with publication information and statistics of sequencing quality. I collected the relationships between the archived data and the published articles by text mining. The statistics of sequencing quality was calculated by using FastQC and processed with metadata. The developed system integrated the sample metadata with the related publication, which also enabled researchers to find related data from the public database easily. The calculated quality statistics of each sequencing data can provide more comparable attributes of the data rather than free text. The additional information helped to supplement the lack of description in the metadata, which also helps researchers to interpret the output of data analysis.

The description of software used in the data analysis is also crucial to evaluate the output of the analysis. To describe the process of data analysis without any uncertain points, I developed a method to package the operations in an executable form with runtime information. The developed system named CWL-metrics works with the workflows described in the Common Workflow Language (CWL), a community standard of workflow description. Using CWL, researchers can describe the operations of tools and workflows in a structured form in YAML format. CWL also enables researchers to write their workflows in a format that is executable by multiple workflow runner implementations. CWL-metrics supports researchers to accumulate runtime metrics information of their workflows. The runtime metrics information with the workflow description in CWL can provide the information of what the workflows will do and how the researchers run them, which makes workflows highly portable and reproducible in any computational environment.

The additional information to input data and the method to describe data

analysis workflow in a reproducible form enable researchers to perform data analysis with a precise description of its processes. To demonstrate the data analysis with these methods, I developed a database and a web application using the public HTSeq data. The database, ChIP-Atlas, is to provide the results of data analysis of the public ChIP-Seq and DNase-Seq to show the comprehensive data of transcription factor binding sites and open chromatin regions. By using the proposed methods to make data analysis process transparent, the users of ChIP-Atlas can evaluate the result of the analysis with the precise description of sample metadata and the data analysis workflow.

Form8・Separate Sheet
(様式8・別紙)

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏　　名　大田　達郎

Title
論文題目　Eliminating uncertainty in the process of knowledge acquisition from the large-scale genomic data

申請者は神戸大学農学部において糸状菌の自家不和合性で学士号を取得し、東京大学大学院農学生命科学研究科修士課程において麦類萎縮ウィルスの研究で修士号を取得した。その後2011年にライフサイエンス統合データベースセンター（DBCLS）に研究員として就職し、約7年半、バイオインフォマティクス、とりわけ塩基配列データ解析技術および解析システムの開発を続けてきた。2014年よりDBCLSの一部が国立遺伝学研究所（遺伝研）のキャンパスに移動したため、それにあわせて遺伝研の職員とも連携しつつ研究開発を実施してきた。

論文内容は大きく3つのテーマで構成される。一つは次世代シーケンサー配列のメタデータを論文や配列の精度情報で補完するもの（2章に相当）、一つはデータ処理のパイプラインをコンテナ化して再利用可能とし、それらの性能を比較評価するもの（3章に相当）、そしてChIP-Atlasと名付けられたChIP-SeqおよびDNAse-Seqデータを俯瞰できるウェブサーバ（4章に相当）である。

配列メタデータを補完するアプローチでは、論文と配列情報の紐づけが難しいことを示しつつ、申請者による補完方法の開発について報告している。成果の一例としてSRAデータベースの中身の分析が紹介され、トランスクリプトームやメタゲノム等、配列を取得する実験的アプローチに応じて配列の精度が変化し、その分布が異なることを定量的、視覚的に示せていた。また、解析パイプラインをコンテナ化するアプローチでは、同一のデータに対して異なるパイプラインを効率的に適用し、その性能比較が容易になったことを実証した。実際に申請者が行った性能比較では、クラウドサーバ上におけるデータ処理時間が、次世代シーケンシングにおける配列リード長に依存して変化する様子を明らかにしてみせた。最後のChIP-Atlasサーバでは、7万におよぶ実験データを再解析し、抗原クラスや細胞のタイプによって簡単にデータを探せるインターフェースを構築した。以上の成果はいずれも、大量のDNA配列を効率よく再利用する、あるいはデータを効率よく検索するのに役立つツール群であり、その需要は大きい。実際にChIP-Atlasに関する論文は発表から半年あまりで50回以上引用され、昨年末にEMBO Reports誌に採択されている。申請者の研究は大量データ時代の生命科学における情報の理解法を先取りしたものといえよう。また、申請者がこれまでの研究で主題の一つとしてきた再現性の担保は、バイオインフォマティクスのみならず生命科学全体で問題視されている課題であり、今後の研究の発展も期待される。

申請者は2013年以降、現在までに8報の関連論文（全て英語、査読あり）を著している。そのうち2017年におけるSRAデータベースの品質評価の論文(GigaScience)と、2018

年における桜メタゲノムプロジェクトの論文（Journal of Plant Research)の 2 報において筆頭著者を務めている。これらの業績からも、生命科学を発展・加速させるツールやサーバを構築できる資質や能力を有することは明白である。

以上の事実を総合的に鑑み、審査委員会は本論文が学位の授与に値すると判断した。